SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models

Year: 2023 | *Citations: 270* | *Authors: Ziyi Lin, Chris Liu, Renrui Zhang*

Abstract

We present SPHINX, a versatile multi-modal large language model (MLLM) with a joint mixing of model weights, tuning tasks, and visual embeddings. First, for stronger vision-language alignment, we unfreeze the large language model (LLM) during pre-training, and introduce a weight mix strategy between LLMs trained by real-world and synthetic data. By directly integrating the weights from two domains, the mixed LLM can efficiently incorporate diverse semantics with favorable robustness. Then, to enable multi-purpose capabilities, we mix a variety of tasks for joint visual instruction tuning, and design task-specific instructions to avoid inter-task conflict. In addition to the basic visual question answering, we include more challenging tasks such as region-level understanding, caption grounding, document layout detection, and human pose estimation, contributing to mutual enhancement over different scenarios. Additionally, we propose to extract comprehensive visual embeddings from various network architectures, pre-training paradigms, and information granularity, providing language models with more robust image representations. Based on our proposed joint mixing, SPHINX exhibits superior multi-modal understanding capabilities on a wide range of applications. On top of this, we further propose an efficient strategy aiming to better capture fine-grained appearances of high-resolution images. With a mixing of different scales and high-resolution sub-images, SPHINX attains exceptional visual parsing and reasoning performance on existing evaluation benchmarks. We hope our work may cast a light on the exploration of joint mixing in future MLLM research. Code is released at https://github.com/Alpha-VLLM/LLaMA2-Accessory.